Machine learning for statistical NLP: Advanced

Final course project report

Anni Nieminen

News article simplifier

Fine-tuned sequence-to-sequence model aimed for L2-learners of Swedish

Introduction and motivation

Simplified versions of text are a great learning tool for L2-learners, children, people with cognitive problems, and non-expert readers. However, a lot of the research focuses on sentence-level simplification (Sun, Jin, & Wan, 2021). **In this project, we focus on document-level simplification, with the goal of fine-tuning a transformer model for the task of simplifying news articles. This tool will be aimed for L2-learners of Swedish.** The goal of this model is to produce simplified (and shorter) versions of the inputs - regular news articles in Swedish. Our model will go through two rounds of fine-tuning.

Original article	Firefighters or firemen are people whose job is to put out fires and rescue people. Besides fires, firefighters rescue people and animals from car wrecks, collapsed buildings, stuck elevators and many other emergencies. Firefighting is a job which requires bravery, strength, quick thinking and a wide range of skills. Firefighters are based at a building called a " fire station " (also known as a " firehouse " or " fire hall "). When their help is needed, they drive a vehicle called a " fire engine " or " fire truck " to the scene responding code 1 code 2 or code 3. These vehicles can pump water and foam to put out fires. Fire engines also carry ladders, cutting tools and lots of different types of rescue equipment. Most carry first aid kits to help people who are injured or hurt.
Document-level simplification	**The job of** a firefighter is to put out fires and **save lives from many emergencies**. They are based at a building called a " fire station ". They drive a vehicle called a " fire engine " or " fire truck " to the scene. The vehicle carries many types of rescue equipment **to help people in danger**.
Text summarization	Firefighters or firemen are people whose job is to put out fires and rescue people and animals from many emergencies. Firefighters are based at a building called a " fire station ". When their help is needed, they drive a vehicle called a " fire engine " or " fire truck " which may carry different types of rescue equipment to help people who are injured or hurt to the scene .

Table 1: Examples for document-level simplification and text summarization. It can be seen from the **Bold** part that the simplified article not only deletes complicated and unimportant sentences from the original article but rewrites the clause, merges two sentences then simplifies them, replaces difficult words, etc. These are operations that text summarization does not require.

Above is a great example from the paper by Sun, Jin, and Wan (2021) illustrating the difference between the task of summarization and simplification on document-level. Not only are simplified documents often shorter than the source documents, they contain simplified vocabulary and have shorter sentence structure. The use of only main clauses is common and some concepts might get explained before they are discussed.

For immigrants, reading simplified news articles can be a useful tool for both foreign language learning as well as for the whole integration process. Living as an immigrant here in Sweden, we have personally benefited a lot from both various online news article sources as well as the TV news programs, where news articles are presented in ***lätt svenska*** (easy Swedish.) We find that these news sources producing up-to-date news articles in simplified language allow immigrants not only to learn the language by familiarizing them with common vocabulary but also keep themselves updated on what is going on in the country they live in.

In this project, we aim to see if with two rounds of fine-tuning we can reach somewhat satisfactory results in document-level simplification of Swedish texts.

Choice of model and data collection

Our plan for the project was to start with data collection, and then proceed to fine-tune a multilingual T5-model (mT5) from Hugging Face. We chose to work with a T5 model since they are especially aimed for sequence-to-sequence tasks such as translation and summarization, and our project can be subcategorized under the task of summarization. Choosing a model that has been pre-trained with multilingual data was important as my project specifically deals with Swedish data.

Data collection was one of the most time-consuming and labor-intensive parts of this project, due to our choice of collecting the news article pairs manually as well as inspecting their quality qualitatively. Even though we managed to find some news article corpora online (for instance, Språkbanken has a dataset holding 468 502 sentences extracted from 8 Sidor's articles, available here: https://spraakbanken.gu.se/resurser/attasidor), we were not able to find a ready dataset consisting of regular-simplified article pairs. Implementing web scraping with automated 'matching' was not a valid option due to the time restraints of this project.

The data collection 'routine' we followed consisted of first browsing through the news articles published on 8 Sidor's website, and then searching a news article on the same topic from other Swedish news sources (mainly SVT Nyheter, but also other online news sources such as Aftonbladet and Svenska Dagblad). Once we found a matching article, we read it through to make sure the article bodies match in terms of the content. This was because most of the 'regular' news articles on various topics usually go into more detail regarding the topic or are formatted as a 'live reporting', where information gets added under specific subheadings. In other words, as 8 Sidor's news articles are really simple in nature, we had to make sure that the regular news articles cover at least the same "base" information (answering at least ***What, Where,*** and/or ***Why***). Below is an example of an article pair that was included in our dataset (data point n. 16).

En man i Japan som jobbar med ris. Foto: AP/TT

Brist på ris i Japan

Lyssna

Det är brist på ris
i landet Japan.
I många affärer
är det helt slut på ris.

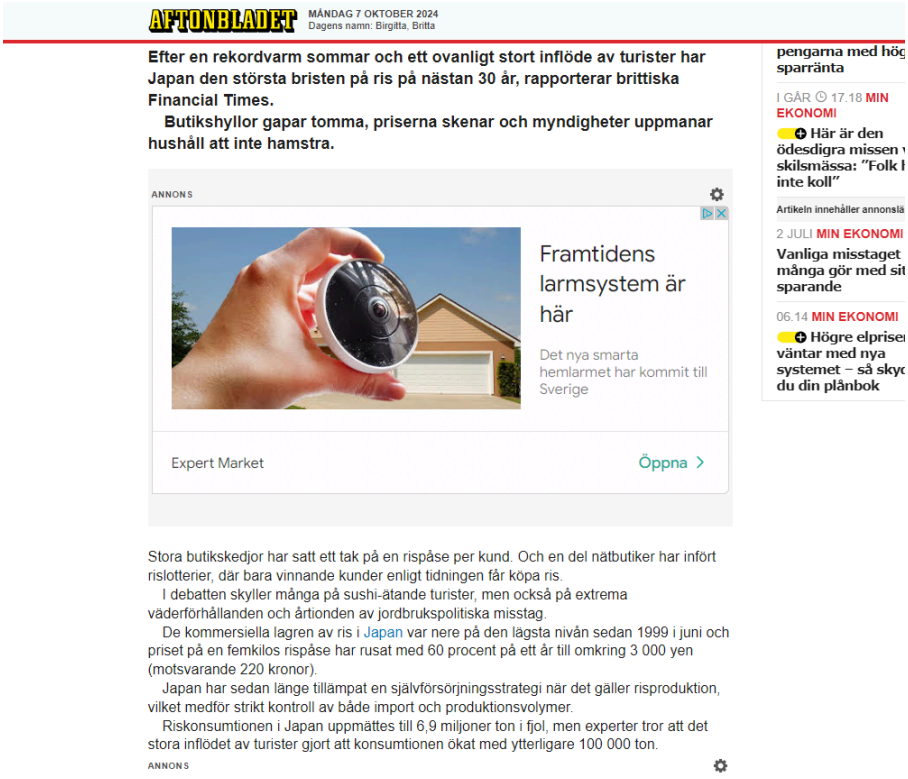
I affärer som har ris kvar
får folk ofta bara köpa
en påse var.

Andra affärer gör ett lotteri
om vem som ska få
köpa riset som finns.

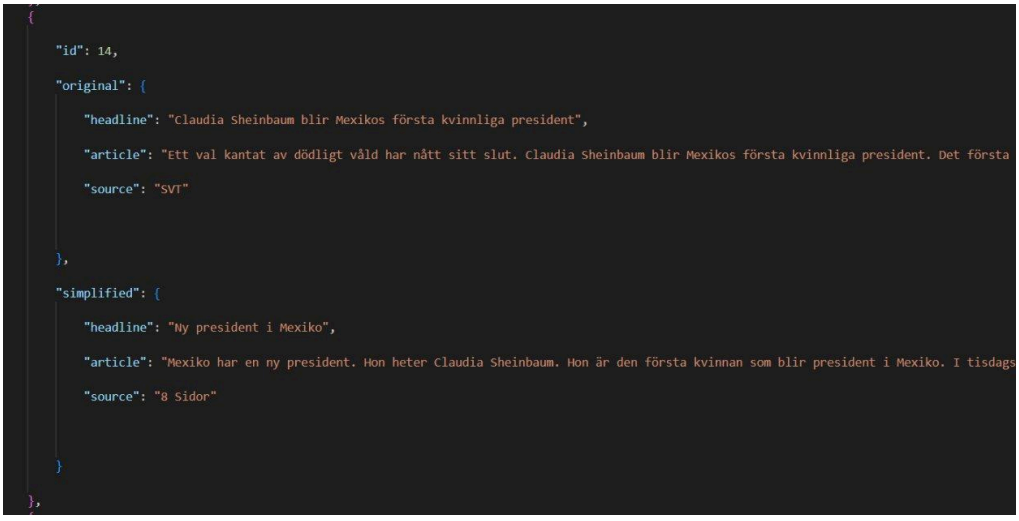
Priset på ris har nu
blivit mycket högre.

Screenshot from 8 Sidor's version on Japan's rice shortage. The article presents the topic on a very basic level. "*I många affärer är det helt slut på ris",* translates to "Rice is completely finished from many stores." "[...] får folk ofta bara köpa en påse var." translates to "people are allowed to only buy one bag".

From this example, we can see that the sentence structure is simple, the paragraphs are clearly separated and the vocabulary is relatively simple.



Screenshot from Aftonbladet's news article on the rice shortage in Japan. Here, the same information is conveyed but the choice of vocabulary is much more advanced. For instance, the information about the limit of purchase is stated as follows: "*Stora butikskedjor har satt ett tak på en rispåse per kund.*"



Above is a screenshot from the json-file, where the collected news article pairs were formatted. The values for the key "article", the bodies of the articles, were used.

Introduction to fine-tuning process

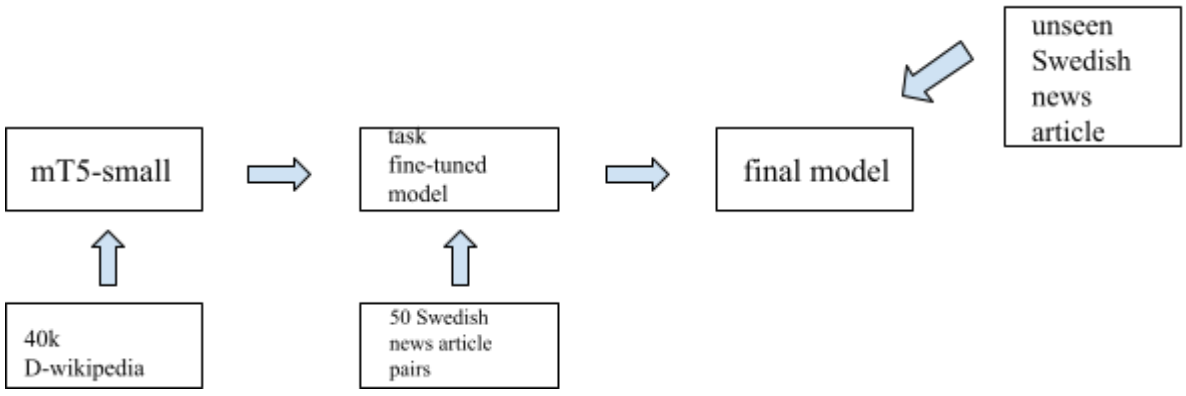
When it comes to the model, we opted for the small version of Google's mT5 (300M parameters), hopeful that the performance of this smaller model would yield satisfactory results. We also came to notice that the fine-tuning process was a lot more time-consuming than I'd expected, regardless of using GPUs.

In our very first attempts, we tried fine-tuning the multilingual mT5 model with our 50 news article pairs. However, these attempts failed miserably, as expected, as the model was only producing the `<extra_id_0>` token along with one other Swedish word. It was obvious that a lot more data would be needed.

This led to us adopting a new approach. We came to the conclusion that it might be a better idea to first fine-tune the model for the down-stream task of simplification first with English data. Consequently, we proceeded to explore fine-tuning the mT5-model with **D-Wikipedia dataset** (Sun, Jin, & Wan, 2021), consisting of 143,546 Wikipedia articles aligned with their simplified versions (many common Wikipedia articles have "simple English" in the language options). Below is an example of a data pair (first one regular Wikipedia article, the second one the Simple English Wikipedia article):

```
annin@anninAsus:/mnt/c/Users/annin/MACHINE_LEARNING_ADVANCED$ cat train.src | head -1
april is the fourth month of the year in the gregorian calendar , the fifth in the early julian , the first of four months to h
ave a length of 30 days , and the second of five months to have a length of less than 31 days.april is commonly associated with
 the season of autumn in parts of the southern hemisphere , and spring in parts of the northern hemisphere , where it is the se
asonal equivalent to october in the southern hemisphere and vice versa .
annin@anninAsus:/mnt/c/Users/annin/MACHINE_LEARNING_ADVANCED$ cat train.tgt | head -1
april is the fourth month of the year , and comes between march and may . it is one of four months to have 30 days.april always
 begins on the same day of week as july , and additionally , january in leap years . april always ends on the same day of the w
eek as december.april 's flowers are the sweet pea and daisy . its birthstone is the diamond . the meaning of the diamond is in
nocence .
```

After this, we would **re-fine-tune** the resulting model with our Swedish news article pairs, with the goal that the model learns to adjust the task of simplification to the Swedish news articles. Below is an illustration of this fine-tuning 'pipeline'.



In the next two chapters, we will discuss these two fine-tuning processes, training results and the models' performances. We will also introduce the SARI metric, which we used for the evaluation.

1st model fine-tuning (for simplification task)

As the first step of the project's fine-tuning 'pipeline', we fine-tuned the **mT5**-model with the D-Wikipedia dataset. Due to time constraints and heavy memory usage on the GPU, the training of the model was done with only 40k data pairs from the dataset. Furthermore, LoRa was used to cut down on the training time. Below is a snippet of the training results after 4 epochs through the training data, which took around 9 hours.

Epoch	Training Loss	Validation Loss	Sari
1	3.741000	2.818675	{'sari': 37.15475633080281}
2	3.452000	2.695584	{'sari': 38.797998502753465}
3	3.332800	2.656935	{'sari': 39.71900946529877}
4	3.287000	2.613796	{'sari': 40.07296562727385}

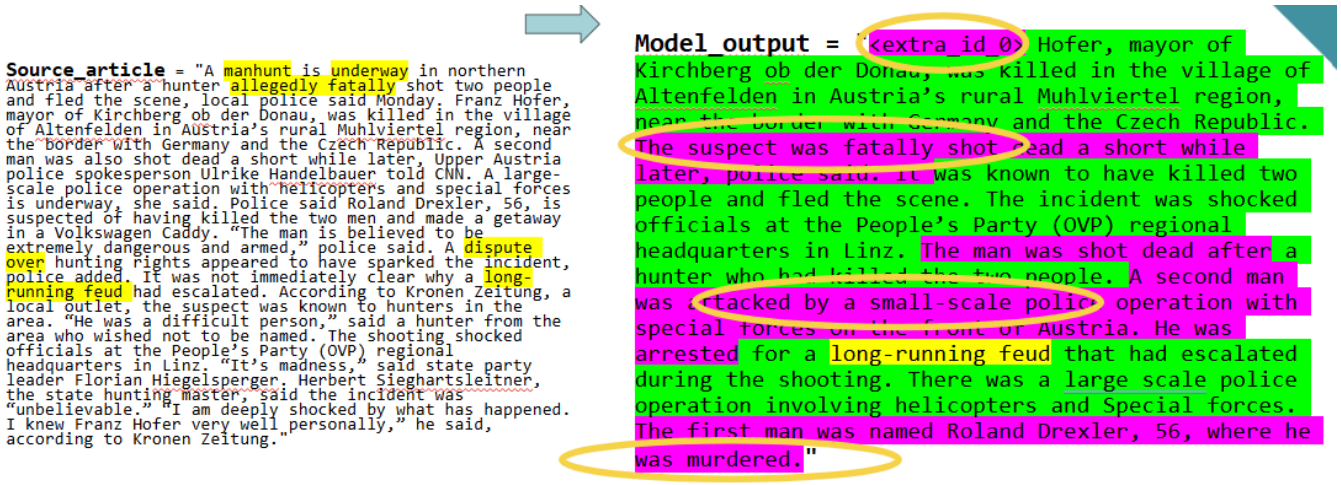
The last metric, SARI, which stands for system output against references and against the input sentence, "is a metric used for evaluating automatic text simplification systems" (https://huggingface.co/spaces/evaluate-metric/sari, accessed November 2nd, 2024). SARI compares the simplified sentences generated by the model against the reference simplifications as well as the inputs. The formula is `( F1_add + F1_keep + P_del) / 3` . The F1-score for add-operations is calculated by examining the simplification predictions made by the model: how often they contain those added (*simplifying*) words that are present in the reference simplified sentences but not in the inputs? F1-keep score looks at how often the predictions contain words that are present both in the inputs and in the reference simplifications. Finally, P-del stands for the precision score of deletion operations - how often are words that have been deleted from the reference sentences also deleted from the model's predictions?

We must take the above sari scores with a grain of salt. This is due to two main reasons: first of all, SARI scores are usually computed against multiple references. In our project, we only had one reference per input. Secondly, SARI's ability to measure meaning preservation is also questionable (Beauchemin, Saggion, and Khoury, 2023).

Next, we will proceed to qualitatively examine the model's performance when it's tested with unseen data. Here are the generation parameters that were used:

```python
from transformers import AutoModelForSeq2SeqLM

model = AutoModelForSeq2SeqLM.from_pretrained("./final-task-fine-tuned-model-40k-traindata/checkpoint-4000")
generated_ids = model.generate(
    inputs,
    do_sample=False,
    repetition_penalty=2.0,
    max_new_tokens=300,
    min_new_tokens=200,   # min length for generation
    num_beams=5,     # n beam search
    no_repeat_ngram_size=3,  # prevent repetitive n-grams
)
```



The words that have been overlined with yellow in the source article are words that we would have liked the model to simplify (i.e. words that foreign language learners might struggle with). The output (on the right) has the following color coding: green for parts that are consistent with the source article, pink for inconsistencies. We can see that even though most of the output seems to factually align with the information in the input, there are many inconsistencies, indicating that the meaning preservation ability of the model is poor. For instance, suspect and victim get confused multiple times in the generated output. Some of the main information is also not present in the output (in this example, the information concerning the ongoing manhunt is lost). The output is also rather repetitive, and most importantly, does not seem to contain simplified vocabulary.

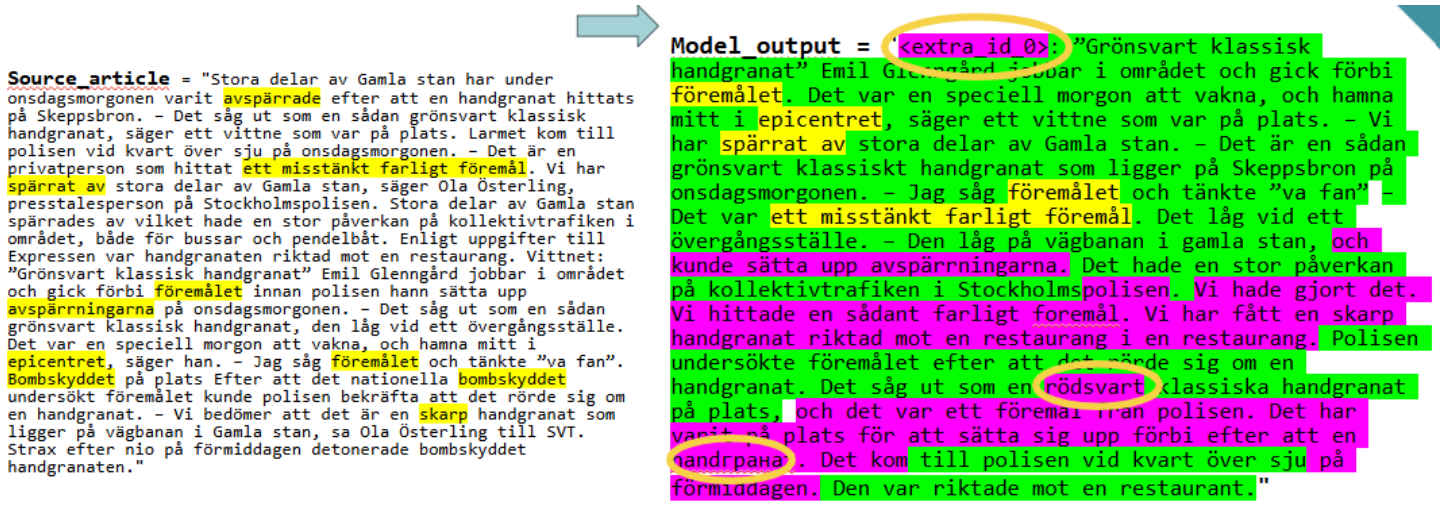
2nd model fine-tuning (for Swedish language)

Regardless of the quite poor performance of the task fine-based model, due to time restrictions we had to proceed to re-fine-tune the fine-tuned model from our previous step with the 50 Swedish news article pairs we had collected. Below is a snippet of the training results after 20 epochs (note that the number of epochs on the screenshot do not match this) through the training data, the 50 Swedish news article pairs.

8	No log	3.051745	{'sari': 33.282336954903556}
10	No log	3.007188	{'sari': 35.10583314679875}
12	No log	2.982897	{'sari': 35.55144491812662}
13	No log	2.976548	{'sari': 35.77589931524155}

Again, we must interpret these SARI results with caution, as we only have one reference simplification per input in our project.

Next, we will qualitatively analyze the performance of the model when we feed it unseen Swedish news articles. The generation parameters were the same as the ones used for testing the task fine-tuned model in the previous paragraph.



The same color codes are used in this example as in the one above. We can establish that, again, the output contains a lot of factual inconsistencies compared to the source article. For instance, in the first example the information concerning the place where they found the hand grenade changes from “vägbanan” (correct) to “restaurang” (incorrect). Also, the color of the grenade suddenly changes from “grönsvart” (correct) to “rödsvart” (incorrect). When it comes to the second example, the numerical information concerning the taxation of snus in the model’s output is inconsistent. The different parties discussed also get mixed up. Additionally, unfortunately, most of the words we were hoping the model would simplify (marked with yellow in the source article) are still present in the simplification, meaning that the model has not really learnt the simplification task. There is also some unnecessary repetition in the output. Furthermore, the model also produces some cyrillic letters (last yellow circulation in the first example)!

However, with our visual quality assessment, it seems like the model's output is mostly correct Swedish, meaning that there aren't many severe grammatical mistakes. Additionally, the sentences are possibly a little bit shorter than in the source article.

Challenges and future work

During our project, we encountered many challenges and limitations.

When fine-tuning transformer models with hundreds of millions of parameters, there are often GPU memory limitations to consider. This was also one of the main challenges in our project.

There are multiple aspects that can be looked into and explored in order to improve the model's performance. First of all, we have to note that the amount of data in the second fine-tuning was not sufficient. To solve this issue, we could try two options: firstly, we could try translating the D-wikipedia data that was used for the first fine-tuning in Swedish. Secondly, we could look into various web-scraping techniques to automate the data collection or even generate more news article pairs synthetically. A more deep qualitative analysis on the D-wikipedia dataset could also prove to be useful.

Furthermore, exploring the performance of models that have been specifically trained with only Swedish data (like GPT-SW3) could be interesting. However, limitations having to do with GPU memory and training time would still be a relevant issue.

Lastly, we must also note that the evaluation metric we used in this project was probably not the most optimal. In the future, we could consider other metrics that take into account the meaning preservation between the input and simplification, such as MeaningBERT (Beauchemin, Saggion, & Khoury, 2023).

Conclusion

In this project, we aimed to fine-tune a mT5-model for the task of document-level simplification. Specifically, we aimed to first fine-tune the model with task-related data, and then re-fine-tune the resulting model with Swedish data. Our goal was to reach somewhat satisfactory results when testing the model with unseen Swedish news articles.

From the qualitative analysis of the outputs produced by the model, we could conclude that the model did not perform well when it comes to the task of simplification. A lot of the words that can be challenging for L2-learners of Swedish could still be found from the model outputs. The model did, however, perform somewhat OK in the task of summarization.

Further work on the project is necessary to improve the results. We believe more work on data collection or synthetic data generation as well as looking into different models are the best ways to proceed.

Bibliography and sources:

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-Level Text Simplification: Dataset, Criteria and Baseline. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu, Wei & Napoles, Courtney & Pavlick, Ellie & Chen, Quanze & Callison-Burch, Chris. (2016). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics. 4. 401-415.

Beauchemin D, Saggion H, Khoury R. MeaningBERT: assessing meaning preservation between sentences. Front Artif Intell. 2023 Sep 22;6:1223924. doi: 10.3389/frai.2023.1223924. PMID: 37808622; PMCID: PMC10557945.

https://github.com/RLSNLP/Document-level-text-simplification/tree/main/Dataset

https://huggingface.co/docs/diffusers/training/lora

https://huggingface.co/papers/2106.09685

https://huggingface.co/spaces/evaluate-metric/sari, accessed November 2nd, 2024

https://8sidor.se/varlden/2024/09/brist-pa-ris-i-japan/, accessed 7th October, 2024

https://www.aftonbladet.se/minekonomi/a/jQO1B0/risbrist-i-japan-lottning-avgor-ratt-att-kopa, accessed 7th October, 2024

https://edition.cnn.com/2024/10/28/europe/austria-mayor-franz-hofer-shooting-intl/index.html, accessed 30th October, 2024

https://www.svt.se/nyheter/lokalt/stockholm/stora-delar-av-gamla-stan-avsparrat-misstankt-farligt-forema, accessed 30th Oct